MODUL 12

REGRESI LINIER SEDERHANA

12.1. Tujuan

1. Mahasiswa mampu menggunakan metode regresi linier.
2. Mahasiswa mampu melakukan analisis regresi linier.
3. Mahasiswa mampu menerapkan metode regresi linier dalam kasus nyata.

12.2. Landasan Teori

Regresi linier adalah metode statistika yang digunakan untuk membentuk model hubungan antara variabel terikat (dependen; Y) dengan satu atau lebih variabel bebas (independent; X). Apabila banyaknya variabel bebas hanya ada satu, disebut sebagai regresi linier sederhana, sedangkan apabila terdapat lebih dari 1 variabel bebas, disebut sebagai regresi linier berganda.

Regresi linear sederhana ataupun regresi linier berganda pada intinya memiliki beberapa tujuan, yaitu :

1. Menghitung nilai estimasi rata-rata dan nilai variabel terikat berdasarkan pada nilai variabel bebas.
2. Menguji hipotesis karakteristik dependensi.
3. Meramalkan nilai rata-rata variabel bebas dengan didasarkan pada nilai variabel bebas diluar jangkauan sampel.

Di dalam suatu model regresi akan ditemukan koefisien-koefisien. Koefisien pada model regresi sebenarnya adalah nilai

©Muhammadiyah university Press

duga parameter di dalam model regresi untuk kondisi yang sebenarnya (true condition), sama halnya dengan statistik mean (rata-rata) pada konsep statistika dasar. Hanya saja, koefisien-koefisien untuk model regresi merupakan suatu nilai rata-rata yang berpeluang terjadi pada variabel Y (variabel terikat) bila suatu nilai X (variabel bebas) diberikan.

Pada analisis regresi sederhana, ada beberapa asumsi dan persyaratan yang perlu diperiksa dan diuji, beberapa diantaranya adalah :

1. Variabel bebas tidak berkorelasi dengan disturbance term (Error). Nilai disturbance term sebesar 0 atau dengan simbol sebagai berikut: E(U/X)=0.
2. Jika variabel bebas lebih dari satu, maka antara variabel bebas (explanatory) tidak ada hubungan linier yang nyata.
3. Model regresi dikatakan layak jika angka signifikansi ANOVA sebesar < 0.05.
4. Variabel independen (predictor) yang digunakan sebagai variabel bebas harus layak. Kelayakan ini diketahui jika angka Standard Error of Estimate < Standard Deviation.
5. Koefisien regresi harus signifikan. Pengujian dilakukan dengan Uji T. Koefisien regresi signifikan jika T hitung > T table (nilai kritis). Nilai T Table, F-Table dan lain-lain bisa dilihat dalam file **StatisticTable.pdf** yang bisa anda unduh di http://yusufsn.staff.ums.ac.id/my-files dalam folder Praktikum Data Mining.
6. Model regresi dapat diterangkan dengan menggunakan nilai koefisien determinasi (KD = r^2 x 100%) semakin besar nilai tersebut maka model semakin baik. Jika nilai mendekati 1 maka model regresi semakin baik.
7. Data harus berdistribusi normal.
8. Data berskala interval atau rasio.
9. Kedua variabel bersifat dependen, artinya satu variabel merupakan variabel bebas sedang variabel lainnya variabel terikat.

©Muhammadiyah university Press

Prediksi terhadap suatu variabel Y (*dependent / response*) menggunakan metode regresi linier didasarkan pada nilai-nilai dari variabel lainnya, X (*independent / predictor*) serta hubungan antar dua variabel.

12.3. Alat dan Bahan

1. Komputer dengan sistem operasi Windows.
2. Program aplikasi RapidMiner.
3. Modul Praktikum Data Warehousing dan Data Mining.

12.4. Langkah-langkah Praktikum

Contoh Kasus :

Dalam sebuah kelas yang memiliki 10 siswa dilakukan sebuah survei terhadap lama belajar seorang siswa dan nilai hasil ujiannya. Data siswa tersebut akan kita gunakan sebagai dasar perhitungan untuk memprediksi nilai ujian terhadap siswa lain berdasarkan lama belajarnya.

12.4.1. Mencari Nilai t-hitung dan Model Regresi Linier

Hipotesis :

Bagaimana mencari nilai t-hitung dan model regresi linier dari data siswa tersebut menggunakan RapidMiner.

Berikut Tabel Data Siswa :

©Muhammadiyah University Press

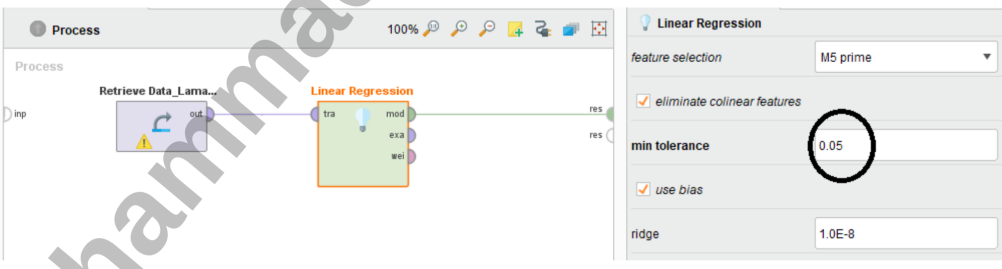
NO_SISWA	NAMA	LAMA BELAJAR (JAM)	NILAI
S-101	JOKO	15	783
S-102	AGUS	18	877
S-103	SUSI	7	505
S-104	DYAH	9	860
S-105	WATI	15	968
S-106	IKA	17	793
S-107	EKO	10	752
S-108	YANTO	5	571
S-109	WAWAN	8	667
S-110	MAHMUD	15	723

4. Buka Ms. Excel, dan buatlah tabel data siswa tersebut. Simpan dengan nama **Tabel_LamaBelajardanNilaiUjian.xls** (Format Excel 2003 *.xls).
5. Jalankan aplikasi **RapidMiner**.
6. Gunakan file **Tabel_LamaBelajardanNilaiUjian.xls** sebagai data yang akan digunakan dalam proses Regresi Linier. Import file ini ke dalam repositori (seperti pada Modul 8 Kegiatan 8.4.2). Yang perlu diperhatikan hanya pada saat penentuan tipe atribut seperti langkah 4.
7. Ubah tipe data dan jenis masing-masing atribut sebagai berikut.
   a) NO_SISWA : polynomial, id
   b) NAMA : pilih Exclude column
   c) LAMA JAM BELAJAR : integer
   d) NILAI : integer, label

©Muhammadiyah University Press

	NO_SISWA *polynominal* *id*	NAMA *polynominal*	LAMA JAM BELAJAR *integer*	NILAI *integer* *label*
1	S-101	JOKO	15	783
2	S-102	AGUS	18	877
3	S-103	SUSI	7	505
4	S-104	DYAH	9	860
5	S-105	WATI	15	968
6	S-106	IKA	17	793
7	S-107	EKO	10	752
8	S-108	YANTO	5	571
9	S-109	WAWAN	8	667
10	S-110	MAHMUD	15	723

8. Pada langkah ini, beri nama **Data_LamaBelajardanNilaiUjian** dan masukkan pada repositori, kemudian klik **Finish**.
9. Gunakan **Data_LamaBelajardanNilaiUjian** ini dan masukkan ke dalam area process.
10. Tambahkan operator **Modeling → Predictive → Functions → Linear Regression.** Hubungkan port out (output) operator Retrieve Data ke entry tra (training) operator ini dan 3 outputnya ke connector res panel.
11. Klik pada operator **Linear Regression**, tentukan parameter **min tolerance = 0.05** (batas toleransi sebesar 5%).



12. Jika data input bertipe nominal atau polynomial, tambahkan operator **Blending → Atrributes → Types → Nominal to Numerical** tepat setelah data input, sebelum operator **Linear Regression**. Set parameter *coding type* menjadi *unique integer*.

©Muhammadiyah University Press

13. Jalankan proses dengan menekan tombol **Run** (atau menekan tombol F11).
14. Berikut hasil proses regresi linier :
    a) Table View (mencari besarnya nilai t-hitung)

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
LAMA JAM BELAJAR	21.608	7.645	0.707	1	2.827	0.022	**
(Intercept)	492.769	96.909	?	?	5.085	0.001	****

Dari tabel tersebut dapat dilihat nilai t-statistic (t-hitung) sebesar 2,827. Berdasarkan aturan statistik, variabel X dikatakan mempengaruhi secara signifikan terhadap Y jika nilai t-hitung > t-table.

Berikut potongan tabel T:

t-table	min tolerance		
α (1 tail)	0.05	0.025	0.01
α (2 tail)	0.1	0.05	0.02
degree of freedom	Nilai Uji T		
1	6,3138	12,7065	31,8193
2	2,9200	4,3026	6,9646
.....	.....	.....	.....
.....	.....	.....	.....
7	1,8946	2,3646	2,9980
9	1,8331	2,2621	2,8214
10	**1,8124**	2,2282	2,7638

Keterangan:

α (1 tail) = regresi linier sederhana
α (2 tail) = regresi linier berganda
degree of freedom = jumlah sampel yang digunakan

Jika t-hitung= 2,827 sedangkan t-table = 1,812, maka 2,827 > 1,812 dengan nilai toleransi 5% (0,05). Sehingga dapat dikatakan bahwa Lama Belajar (X) mempengaruhi secara signifikan terhadap Nilai Ujian (Y).

©Muhammadiyah University Press

b) Text View (mencari model regresi)

LinearRegression

```
  21.608 * LAMA JAM BELAJAR
+ 492.769
```

Berdasarkan hasil pada Text View terlihat sebuah persamaan berikut:

= 21.608 * LAMA JAM BELAJAR + 492.769

Maka dapat dibuat sebuah model persamaan regresi linier sederhana untuk mencari nilai variabel Y (Nilai Ujian) berdasarkan variabel X_1 (Lama Belajar).

Berikut model regresi linier yang terbentuk:

$$\mathbf{Y = 21{,}608\ X_1 + 492{,}769}$$

Dengan model tersebut, dapat dicari Nilai Ujian (Y) dengan memasukkan nilai Lama Belajar pada variabel X_1.

12.4.2. Mencari Nilai t dan Model Regresi Linier Menggunakan RapidMiner

Hipotesis :

Bagaimana memprediksi nilai ujian siswa berdasarkan lama belajarnya menggunakan model regresi linier yang telah dihasilkan menggunakan RapidMiner.

Berikut Tabel Data Siswa :

©Muhammadiyah university Press

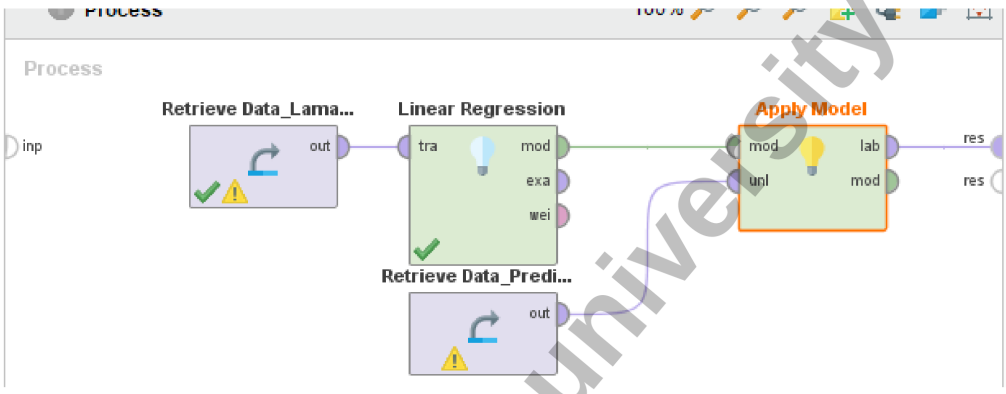
NO_SISWA	NAMA	LAMA BELAJAR (JAM)
S-111	BUDI	12
S-112	SANTI	13
S-113	DIAN	14
S-114	DANI	11
S-115	AHMAD	5
S-116	BAYU	13
S-117	RISA	9
S-118	RANI	10
S-119	YANI	10
S-120	RATIH	9

1. Buka Ms. Excel, dan buatlah tabel data siswa tersebut. Simpan dengan nama **Tabel_PrediksiNilaiUjian.xls** (Format Excel 2003 *.xls).
2. Jalankan aplikasi **RapidMiner**.
3. Gunakan file **Tabel_PrediksiNilaiUjian.xls** sebagai data testing. *Import file* ini ke dalam repositori.
4. Ubah tipe data dan jenis masing-masing atribut sebagai berikut.
   a) NO_SISWA : polynomial, id
   b) NAMA : pilih Exclude column
   c) LAMA JAM BELAJAR : integer

	NO_SISWA polynominal id	NAMA polynominal	LAMA JAM BELAJAR integer
1	S-111	BUDI	12
2	S-112	SANTI	13
3	S-113	DIAN	14
4	S-114	DANI	11
5	S-115	AHMAD	5
6	S-116	BAYU	13
7	S-117	RISA	9
8	S-118	RANI	10
9	S-119	YANI	10
10	S-120	RATIH	9

©Muhammadiyah University Press

5. Pada langkah ini, beri nama **Data_PrediksiNilaiUjian** dan masukkan pada repositori, kemudian klik **Finish**.
6. Tetap menggunakan proses pada kegiatan 12.4.1, masukkan **Data_PrediksiNilaiUjian** ini ke dalam area process.
7. Tambahkan operator **Scoring → Apply Model**, letakkan setelah operator Linear Regression. Hubungkan port-port output dan input seperti gambar berikut.



8. Dengan menggunakan parameter yang sama pada operator **Regression Linear**, jalankan proses dengan menekan tombol **Run** (atau menekan tombol F11).
9. Berikut hasil proses prediksi terhadap data testing menggunakan regresi linier:

©Muhammadiyah University Press

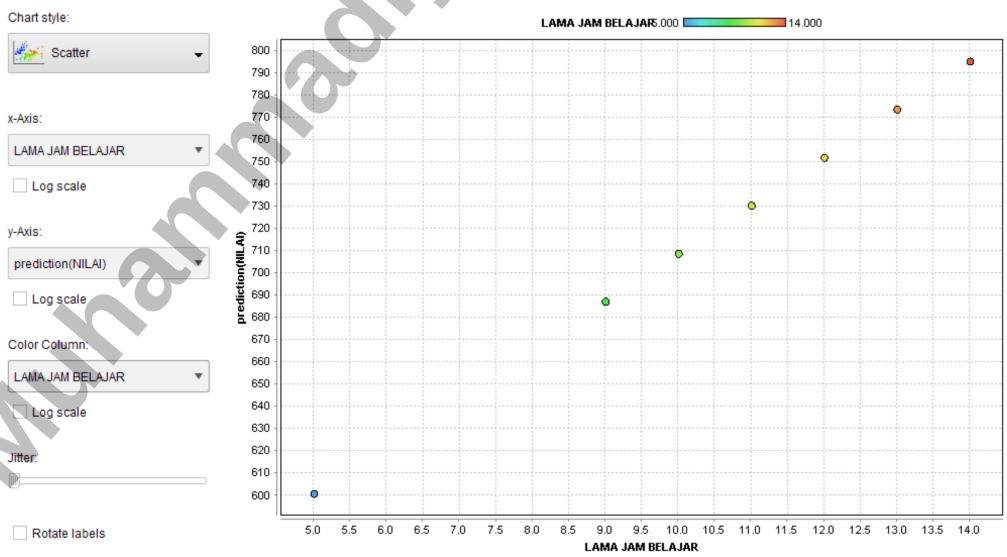
a) Data View (hasil prediksi nilai ujian)

ExampleSet (10 examples, 2 special attributes, 1 regular attribute)

Row No.	NO_SISWA	prediction(NILAI)	LAMA JAM BELAJAR
1	S-111	752.061	12
2	S-112	773.668	13
3	S-113	795.276	14
4	S-114	730.453	11
5	S-115	600.807	5
6	S-116	773.668	13
7	S-117	687.238	9
8	S-118	708.845	10
9	S-119	708.845	10
10	S-120	687.238	9

Hasil prediksi nilai ujian terhadap 10 siswa lainnya.

b) Charts View (Scatter Plot)



©Muhammadiyah University Press

12.4.3. Pembuktian Model Regresi

Pada tahap ini kita akan melakukan pembuktian hasil prediksi menggunakan RapidMiner dengan hasil prediksi menggunakan model regresi yang telah dihasilkan pada kegiatan 12.4.1 berikut.

$$\mathbf{Y = 21{,}608\ X_1 + 492{,}769}$$

E3

	A	B	C	D	E
1	NO_SISWA	NAMA	LAMA BELAJAR (JAM)	Prediction (NILAI)	Prediction (NILAI)
2				Tabel	Model Regresi
3	S-111	BUDI	12	752,061	
4	S-112	SANTI	13	773,668	
5	S-113	DIAN	14	795,276	
6	S-114	DANI	11	730,453	
7	S-115	AHMAD	5	600,807	
8	S-116	BAYU	13	773,668	
9	S-117	RISA	9	687,238	
10	S-118	RANI	10	708,845	
11	S-119	YANI	10	708,845	
12	S-120	RATIH	9	687,238	
13					

Sheet1 Sheet2 Sheet3

Memasukkan nilai variabel X1 ke dalam model regresi (3 contoh data):

1. No. Siswa = S-111, Nama = Budi, memiliki nilai X1 = 12 Jam. Sehingga, Y = (21,608 x 12) + 492,769 = 752,065
2. No. Siswa = S-115, Nama = Ahmad, memiliki nilai X1 = 5 Jam.
   Sehingga, Y = (21,608 x 5) + 492,769 = 600,809
3. No. Siswa = S-120, Nama = Ratih, memiliki nilai X1 = 9 Jam. Sehingga, Y = (21,608 x 9) + 492,769 = 687,241

Berikut tabel hasil pembuktiannya :

Gunakan Ms. Excel untuk melakukan pembuktian. Pada sel **E3** masukkan formula "=(21,608*C3)+492,769". Copy dan Paste pada sel **E4** sampai **E12**.

©Muhammadiyah University Press

E3 =21,608*C3+492,769

	A	B	C	D	E
1	NO_SISWA	NAMA	LAMA BELAJAR (JAM)	Prediction (NILAI)	Prediction (NILAI)
2				Tabel	Model Regresi
3	S-111	BUDI	12	752,061	752,065
4	S-112	SANTI	13	773,668	773,673
5	S-113	DIAN	14	795,276	795,281
6	S-114	DANI	11	730,453	730,457
7	S-115	AHMAD	5	600,807	600,809
8	S-116	BAYU	13	773,668	773,673
9	S-117	RISA	9	687,238	687,241
10	S-118	RANI	10	708,845	708,849
11	S-119	YANI	10	708,845	708,849
12	S-120	RATIH	9	687,238	687,241
13					

Sheet1 Sheet2 Sheet3

Dapat dilihat bahwa hasil prediksi menggunakan RapidMiner menghasilkan nilai yang sama dengan menggunakan Model Persamaan Regresi Linier yang telah dihasilkan pada kegiatan 12.4.1.

12.5. Tugas

Kasus :

Dalam sebuah survei terhadap 15 kepala keluarga telah diperoleh variabel pendapatan rata-rata perbulan, jumlah anggota keluarga yang tinggal serumah, dan daya beli rata-rata perbulan.

Hipotesis :

Bagaimanakah model regresi linier yang terbentuk, dan lakukan prediksi terhadap 10 data yang belum diketahui nilai daya belinya.

1. Buatlah tabel berikut dengan menggunakan Microsoft Excel.

©Muhammadiyah University Press

Tabel Hasil Survei 15 Kepala Keluarga :

NO. RESPONDEN	PENDAPATAN (RUPIAH)	JUMLAH ANGGOTA KELUARGA	DAYA BELI (RUPIAH)
1	1.000.000	6	834.000
2	1.400.000	7	1.200.000
3	200.000	3	134.000
4	1.400.000	6	1.167.000
5	500.000	3	334.000
6	1.700.000	5	1.360.000
7	400.000	3	267.000
8	1.900.000	5	1.520.000
9	300.000	3	200.000
10	500.000	4	375.000
11	700.000	7	600.000
12	1.900.000	3	1.267.000
13	800.000	4	600.000
14	1.500.000	4	1.125.000
15	1.300.000	7	1.115.000

2. Buatlah proses Regresi Linier Sederhana menggunakan RapidMiner dengan ketentuan sebagai berikut.
   a) Variabel bebas (X) = Pendapatan (X1), Jumlah Anggota Keluarga (X2)
   b) Variabel terikat (Y) = Daya Beli
   c) Toleransi yang digunakan = 5%
3. Tentukan apakah variabel X_1 dan X_2 mempengaruhi secara signifikan terhadap nilai variabel Y berdasarkan besarnya nilai t-stat?
4. Tuliskan model persamaan regresi linier sederhana yang terbentuk!
5. Gunakan data testing untuk menjawab perintah berikut:
   a) Lakukan prediksi Daya Beli (Y) dengan menggunakan Model Persamaan Regresi Linier dari hasil pertanyaan nomor 4!

©Muhammadiyah University Press

b) Lakukan prediksi Daya Beli (Y) menggunakan RapidMiner!

Data Testing Yang Digunakan Untuk Prediksi :

NO. RESPONDEN	PENDAPATAN (RUPIAH)	JUMLAH ANGGOTA KELUARGA
1	900.000	5
2	800.000	3
3	500.000	2
4	1.900.000	6
5	600.000	2
6	800.000	5
7	1.000.000	6
8	1.100.000	4
9	1.000.000	4
10	500.000	3

6. Gambarkan pola sebaran data menggunakan Plot View (Scatter) dengan ketentuan berikut :
   a) x-Axis = Pendapatan (Rupiah),
      y-Axis = Prediction (Daya Beli (Rupiah)),
      Color Column = Prediction (Daya Beli (Rupiah))
   b) x-Axis = Jumlah Anggota Keluarga,
      y-Axis = Prediction (Daya Beli (Rupiah)),
      Color Column = Prediction (Daya Beli (Rupiah))

©Muhammadiyah University Press